摘要

小米作为国内知名的传统手机厂商，近年来突然宣布迈入车企行业，其首辆汽车小米 su7 在近期召开上市发布会，引发了大量的网络舆情讨论。为探究社交网络对该事件的舆情聚焦点，以及公众对其的情感倾向态度，帮助小米企业掌握舆情风波，为后续研发提供可鉴的对策建议，本文基于主题归纳、深度学习方法建立相关模型，对事件舆情进行探究。

本文通过网络爬虫技术，爬取目前视频社交媒体上对小米 su7 的用户评论数据，共计 63739 条，作为本文研究的数据来源。依据网络舆情热度演化趋势，将整个舆情划分为预热期、爆发期、衰退期三个阶段，基于 LDA 主题建模对不同阶段下的文本含义进行特征提取，并基于文本特征对不同阶段的舆情聚焦点进行总结概括；此外，为最大程度上得到准确的情感分类结果，本文基于人工标注与伪标签法结合的方式，构建了一个在小米汽车领域下的特定数据集，利用该数据集基于一种引入注意力机制，且结合 CNN 卷积神经网络特征提取的 Bi-LSTM 模型，对所有用户评论进行情感分类。基于上述分类后的情感数据集，以探究不同阶段下公众对小米 su7 的情绪动态演化趋势，以及各阶段下影响公众情绪特征的聚焦因素。

整理上述研究结果，本文得出如下主要结论：（1）公众的主要舆情聚焦点，在于对小米品牌新进车企行业的消费信任感缺失，该类担忧贯穿了小米汽车舆情的三个发展阶段，各阶段对其的担忧侧重点虽有所转变，但仍主要集中于定价策略、营销手段、性能配置以及外观设计四个方面。（2）公众在舆情中的情感态度，整体持中立，但消极情绪在三个阶段中所持的比例，均高于积极情绪，有相当部分消费者对小米汽车的发售持怀疑态度，其聚焦因素在不同阶段呈现多样性；（3）舆情的最后阶段，公众积极情绪一路上升，直至与消极情绪的比例极为接近，小米汽车的发售随着进入大众视野，其得到了众多消费者的认可，但其舆情仍然呈现出情绪的极端分化。

此外，基于上述舆情探究结论，本文对小米汽车的未来发展提出了相关对策建议。

关键词：小米汽车；舆情阶段化；LDA 主题分析；情感分类；深度学习

Abstract

Xiaomi, a well-known traditional mobile phone manufacturer in China, has recently made headlines by entering the automobile industry. The launch of its first car, the Xiaomi su7, sparked a significant online public sentiment discussion. This study aims to explore the focal points of social media sentiment regarding this event, and the public's emotional tendencies, to help Xiaomi navigate public sentiment and provide insights for future research and development. Employing thematic induction and deep learning methods, models were developed to examine the sentiment surrounding this event.

This paper utilized web scraping techniques to collect user comments on the Xiaomi su7 from video social media platforms, amounting to 63,739 comments as the data source for this study. According to the evolutionary trend of online public sentiment, the sentiment was divided into three phases: pre-launch, burst, and decline. Using LDA topic modeling, text feature extraction was conducted for each phase, and a summary of the public sentiment focus for each stage was developed. Moreover, to achieve accurate sentiment classification, a custom dataset specific to Xiaomi's automotive domain was constructed by combining manual annotation with pseudo-labeling. An attention-enhanced Bi-LSTM model incorporating CNN features was utilized for sentiment classification of user comments. The classified sentiment data set was then used to investigate the dynamic trends of public emotion toward the Xiaomi su7 and the factors influencing public sentiment at different stages.

The study concluded the following key findings: (1) The primary focus of public sentiment was the lack of consumer trust in Xiaomi's new venture into the automotive industry, a concern consistent across all stages, focusing mainly on pricing strategy, marketing tactics, performance specifications, and design. (2) Overall, the public's emotional attitude was neutral, but the proportion of negative sentiments outweighed positive ones in all phases, with considerable skepticism towards the launch of Xiaomi's car, varying focus factors at different stages. (3) In the final stage of sentiment analysis, positive sentiments rose steadily, nearly equaling negative sentiments, as the Xiaomi

car gained recognition among consumers, though the sentiment remained highly polarized.

Based on the findings, this paper provides strategic recommendations for the future development of Xiaomi's automotive endeavors.

Keywords: Xiaomi automotive; Phasic public opinion; LDA topic analysis; Sentiment classification; Deep learning

目录

表格与插图清单

视频社交媒体视角下小米汽车 su7 的网络舆情倾向性分析
——基于主题特征与深度学习方法

一、绪论

（一）研究背景

近年来，中国新能源汽车产业呈现出良好的发展态势，已初步在全球范围内形成了综合竞争优势。2023 年 6 月 2 日，国务院常务会议着重讨论了促进新能源汽车产业高质量发展的相关政策和措施，为新能源汽车的发展提供了政策支持。

据中国汽车工业协会的数据，2023 年 1 月至 9 月，中国新能源汽车行业维持了其强劲的增长趋势，产销量分别达到 631.3 万辆和 627.8 万辆，年增长率分别为 33.7%和 37.5%，新能源车在汽车总销量中的占比提升至 29.8%。新能源汽车已成为中国制造业的一张新名片。

小米作为国内知名的传统手机厂商，其在近年来突然宣布迈入汽车行业，小米 CEO 雷军称“要为小米汽车而战”，一时引发了巨大的网络舆情讨论。一方面，小米作为曾经手机性价比的代表性厂商，不由令广大消费者关注其汽车发售价格，其能否延续性价比的核心策略，将新能源汽车带入大众能够接受的平价视野，尤被消费者重视。

再者，小米汽车加入新能源行业的大胆尝试，无疑加剧了市场竞争，其能否一定程度上影响当下新能源汽车的行业格局，为新能源汽车发展提供更多可借鉴的新思路、新方案，推动整个行业技术进步和产业升级，为消费者提供更多优质的选择，既被广泛消费者群体重视，同样备受业界各新能源企业迫切关注。

此外，目前阶段，中国新能源汽车市场正面临着从“政策导向型市场”逐渐向“市场导向型市场”转型，私人购买新能源汽车的兴起，恰好助力小米 su7 更顺利地走入大众视野。因此本文以小米汽车为切入点，探究其网络舆情的倾向性，既可为小米汽车的未来发展提供消费者视角下的舆情聚焦点，同时能够为整个新能源行业的舆情应对，提供大数据视角下借鉴意义。

（二）文献综述

1.情感分析方法研究现状

现如今，随着越来越多的网民在网络平台上发表自己的观点意见，对各大平台用户带个人情感色彩以及具有倾向性言论进行情感分析的工作也在逐步开展。情感分析（Sentiment Analysis）能有效推断一个人对于特定事件或主题的态度，也可以指从一个人的表达中判断其情绪状态。文本的情感分析任务可划分为粗粒度和细粒度，同时基于文本的不同粒度，又可以分为词语、句子、篇章级。根据社交媒体舆情中的文本情感分析研究一般定义为句子级情感分析任务。再通过相关技术分类不同方面所表达的情感倾向，目前情感分析方法主要分为三类：基于情感词典的情感分析方法、基于传统机器学习的情感分析方法和基于深度学习的情感分析。

①基于情感词典

基于情感词典的情感分析方法，首先是构造出包含情感词的情感词典，然后基于词典的方法来从文本中提取情感，为文本分配正负面情感标签的过程，以此来判断其中的情感倾向[9]。

在国外研究中，Kim[10]等人最早提出了一个自动从在线评论中提取利弊的系统，即具有相反标签的词典用以情感分类，取得初步良好成效。后 Minqing Hu[11]等人通过挖掘和总结客户对产品的评论，识别意见句中的态度构建出新的情感词典，在一定程度上提高了词典判别的准确度。Agarwal[13]等人讨论了一种方法，即根据推特微博网站上公开的推文的情绪对其进行预处理和分类使得在情感分类系统中，观点主体性的概念得到了解释，实现了一大进步。Ahmed[13]等人介绍了一种构建领域相关情感词典的新方法 SentiDomain,旨在从目标域的句子全局表示中学习一组嵌入的情绪聚类，并构建了多语言词典，证明其提升了判别的准确性。Zhang[14]等人使用微博中的表情符号，结合情感词来构建中文情感语料库，确保了语料库的规模和准确性，进一步扩充了情感词典。

国内对于基于情感词典的情感分析方法相较于国外发展较晚。朱艳辉[1]等人以中文词语相似度计算方法为基础,提出了一种根据中文情感词语的情感权值的计算方法,构建出中文基础情感词典，并结合 TF-IDF 特征权值计算方法,对中

文文本情感倾向进行判别,提升了情感分类的效果。

情感词典的质量与完整度直接决定着文本情感分析的准确性。在当下网络舆论新兴用语层出不穷的情况下，构造出完备的情感词典具有很大难度，同时，基于情感词典的分析方法需要考虑情感词典与研究主题的契合度，使得该方法的域泛化性较差。

②基于机器学习

机器学习的方法包含有监督、无监督和半监督三类。无监督方法无需标记样本，而在有监督方法中，提供的数据皆为有标注。许多研究人员利用有监督的机器学习方法进行情感分析，在国外研究方面 Pang[15]等人最早采用三种机器学习方法（朴素贝叶斯、最大熵分类和支持向量机）融入情感分类的识别中，取得较好实验结果。Neethu[16]等人试图使用机器学习方法来分析关于手机、笔记本电脑等电子产品的推特帖子，提出了一个新的特征向量，用于将推文分为正面、负面，并提取人们对产品的看法，最后成功实现目标。Andreevskaia A[17]等人提出了一种新的方法来解决不同领域的系统可移植性问题：一个情感注释系统，该系统由两个具有基于精度的投票加权的分类器组成，它在准确性和重新调用方面显著提高。

在国内研究方面，唐慧丰[3]等人在以 KNN、SVM 等作为不同文本分类的方法，在此基础上分别进行情感分析，情感分类取得较好的效果。王大伟[2]等人在文本分析中应用了 PCA-SVM 算法，对文本词向量进行降维处理，在最大程度上保留原始数据的特征，后使用 logistic、SVM 等算法进行分析,实验证实 PCA-SVM 算法在中文文本情感分析方面具有一定的优势。彭敏[4]等人在进行特征提取时采取改进的卡方统计方法，并运用 SVM 和朴素贝叶斯算法进行分类，提高了情感识别的准确率。

③基于深度学习

深度学习较于机器学习能够获取到数据内部更深层的内涵，同时提高分类的准确性，因此其也逐渐在情感分析领域中被广泛使用。

Mikolov[18]等人引入 Skip-gram 模型进行词向量学习，并提出 softmax（负采样）方法，使得更有效的处理分析较高数量级的数据集。Denil[19]等人基于扩展

的动态卷积神经网络，该网络在句子和文档级别学习卷积滤波器，分层学习捕获和组合与高级语义概念相关的低级词汇特征，提高了分析准确率。Lan[20]等人通过减少参数，用 SOP 任务替换原始 BERT 中的 NSP 任务，提升了模型针对句子级别的语义理解，从而提高了情感分类的准确性。

于国内研究，胡朝举[5]等人融合情感标签，将 LSTM 与 CNN 进行融合来提取文本上下文信息，大大提高了分析准确率和 F 值。陈葛恒[6]利用双向 LSTM 进行前后推算并结合了极性转移，加强需要获取的文本句子前后关联性，对情感分析的准确性提高。Zhilin Yang[21]等人提出了 XLNet，一种广义的自回归预训练方法，相较于 BERT 取得了更好的实验结果。

但整体而言，深度学习在自然语言文本处理领域在国内起步较晚，且由于中文语言的困难性，其在中文语言处理时具有较大的挑战性。

2.社交媒体情感分析研究现状

社交媒体是人们创作、获取信息、交流意见的网络平台。当下，社交媒体平台已在生活各处扮演者重要角色。因此，通过对社交媒体平台中舆论数据的收集和分析可以获取群众的情感倾向以及多种有益信息。

国外的社交媒体情感分析研究通常从社交媒体平台如 Twitter、Facebook 中收集舆论信息数据进行情感分析。Chikersal[22]等人描述了一个推特情绪分析系统，将规则分类器和监督学习相结合，经实验得出规则可以帮助细化 SVM 的预测。Kalchbrenner[23]等人测试 DCNN 在 Twitter 情感分析中的应用，取得较好结果。Chiorrini[24]等人研究了 BERT 模型在 Twitter 数据的情绪分析和情绪识别中的使用，实验表明，该模型在情绪分析和情绪识别方面分别达到了较好的准确度。

而由于中文语义表述的多样性，易造成一词多义、语义混淆的情况，因此，分析中文社交舆论平台的数据信息难度更高。微博、知乎等作为中国偏文本类社交媒体平台，被运用于进行舆论文本情感性分析的场景较多。Zang[25]等人利用大规模的文本语料库和 KGs 来训练增强的语言表示模型，并合并多个词典进行情感分析。王友卫[7]等人在每个用户个人的兴趣特征基础上构建情感词典，再利用支持向量机融合不同模型分析不同用户的情感倾向，取得较好结果。

然而，国内对于视频社交媒体平台上的网络舆论分析则相对欠缺。例如 B 站、抖音等国内主要的视频社交网络平台，其评论文本篇幅普遍较长，包含了更为复杂的情感信息，并且由于表情、图片等非结构化数据的引入，对舆情情感的捕获造成了巨大挑战。韩坤[8]等人通过结合 BERT 与 TextCNN 模型，提出一种融合 BERT 多层次特征的文本分类模型（BERT-MLFF-TextCNN），并对抖音短视频平台中的相关评论文本数据进行情感分析，为研究视频评论文本数据提供了一定借鉴思路，但该类文献数量仍相对不足。

整体而言，当前大部分情感分析任务大多应用于偏文本类的社会媒体平台，对于视频类的情感分析较少，这是目前对于社交媒体情感分析研究现状的欠缺之处。本文将以此问题为切入点，基于 B 站社交平台中有关小米 su7 的热点视频，结合主题特征与深度学习方法，捕获群众对该话题的关注侧重点与情感态度倾向，以此提出相应对策建议。

（三）论文主要工作

本文一共分为六个章节，各章节主要内容如下：

第一章主要介绍研究背景，对国内外研究现状进行总结概括，并对论文主要工作进行阐述。

第二章主要介绍本文研究中所需要的相关技术理论，包括主题分析方法、情感分类方法、网络爬虫技术以及深度学习的相关技术，深度学习方法主要介绍了下文模型融合使用的卷积神经网络 CNN、双向长短期神经网络以及注意力机制。

第三章主要介绍本文数据的来源以及获取方式，并对所得数据的预处理过程进行了详细阐述，

第四章主要在阶段性舆情划分的基础上，基于词云分析与 LDA 主题特征提取对舆情聚焦点进行了探究。

第五章主要采用伪标签法对小米汽车领域的特定数据集进行了构建，并基于 CNN-Bi-LSTM-ATT 模型的搭建，对用户评论数据集进行情感分类，得到的情感分类数据集，在第四章舆情阶段划分的基础上，对舆情进行情绪的动态演化分析。

第六章总结了本文的主要工作，并得出研究结论以及对小米汽车未来发展的对策建议，同时阐述了本研究的不足之处。

本文技术路线图如下图 1 所示：

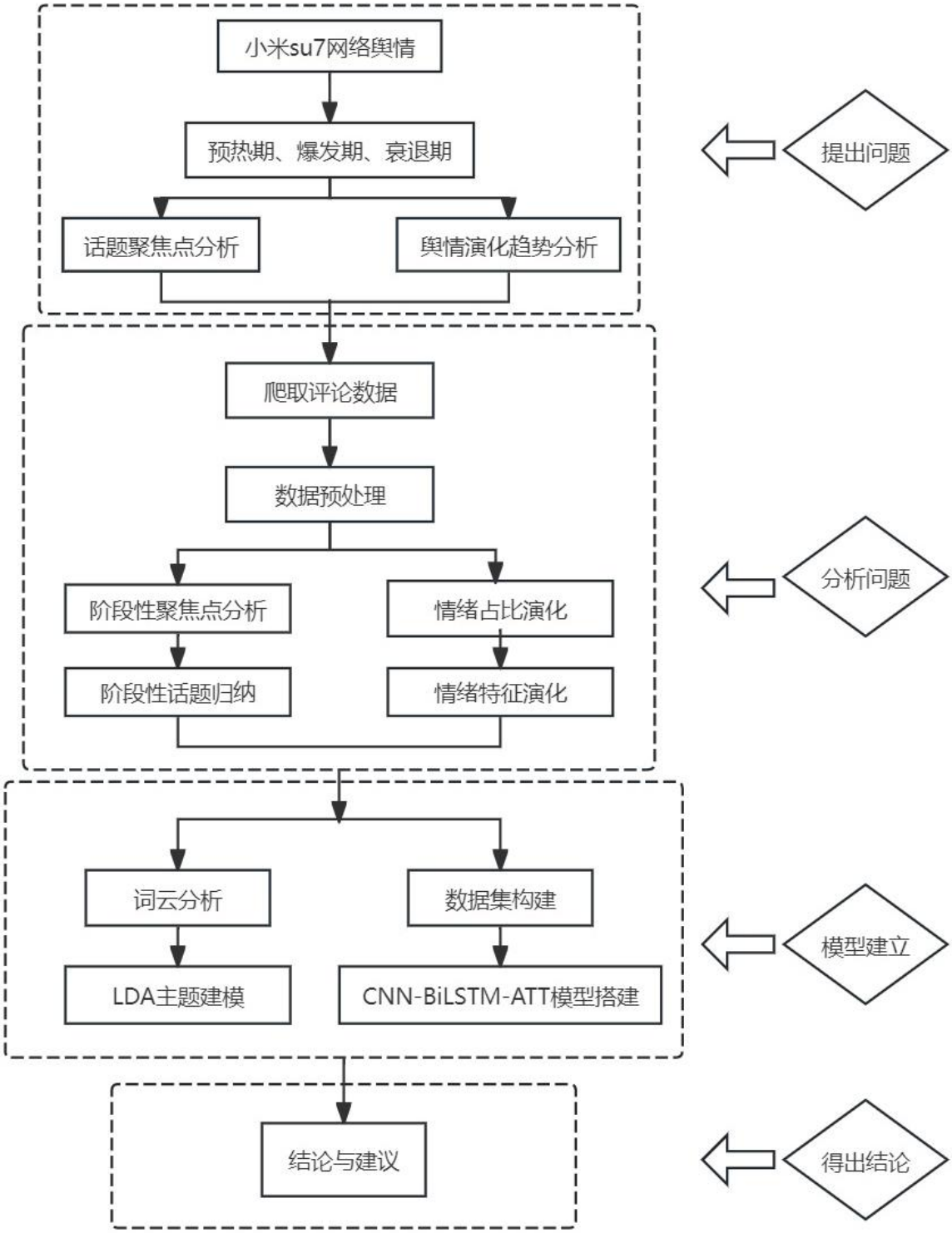

图 1 技术路线图

二、相关理论及技术介绍

（一）主题分析方法介绍

主题分析即从大量的文本数据中识别主要主题内容，帮助用户快速理解文本中的内容与重点，提高搜索引擎的效率，常见的主题分析方法有 TF-IDF、LDA 等，本节中主要对下文使用的 LDA 主题模型进行介绍。

LDA 全称为潜在狄利克雷分配（Latent Dirichlet Allocation），是一种无监督的机器学习技术，主要用于推测文档中的主题分布，将文档集中每篇文档的主题以概率分布的形式表示出根据主题进行主题聚类或文本分类的结果。LDA 模型通常使用词袋特征（bag-of-word feature）来表示文档。

LDA 主题模型基于贝叶斯概率的三级模型，将每个数据集视为一组可能的主题合集，三个层次结构分别为包含词、主题和文档。该模型文档生成方式如下：

(1)从狄利克雷分配 α 中取样生成文档 i 的主题分布 θ_i。

(2)从主题的多项式分布 θ_i中取样生成文档 i 第 j 个词的主题词的主题 $z_{i,j}$。

(3)从狄利克雷分配 β 中取样生成主题$z_{i,j}$对应的词语分布 $\varphi_{zi,j}$。

(4)从词语多项分布 $\varphi_{zi,j}$中采样最终生成词语 $w_{i,j}$。

其中,模型主要示意如下图 1 所示:

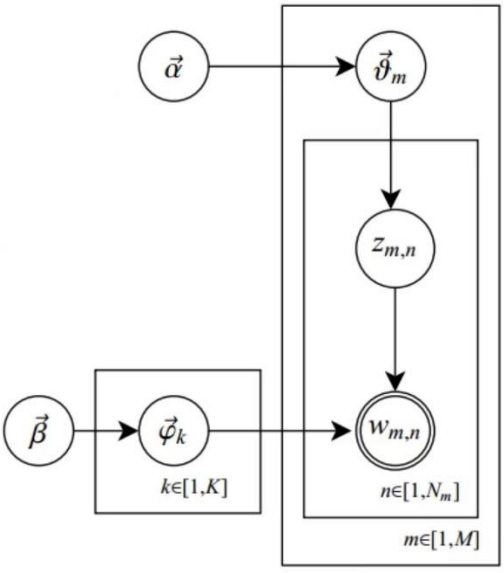

图 2 LDA 主题模型示意图

（二）情感分析方法概述

情感分析也被称为倾向性分析、观点挖掘、情感分类，是自然语言处理、文本挖掘的一种重要的信息分析处理技术。其目的是识别和提取源自文本的主观信息，主要用于理解个人、群体或社会大众对特定主题、产品或服务的情感倾向、情绪状态和观点。目前，于文本的情感分类而言，主要有三种方法，分别是基于情感词典的情感分析方法、基于传统机器学习的情感分析方法和基于深度学习的情感分析。

1.基于情感词典

基于情感词典的情感分析方法依赖于预先定义的情感词典来评估文本中的情感倾向，根据不同情感词典所提供的情感词的情感极性来实现不同粒度下的情感极性划分。

该方法虽然可以准确反映文本的非结构化特征，易于分析和理解，但是由于其主要依赖于情感词典的构建，需要情感词典的不断扩充来满足对新词的准确识别，同时在使用情感词典进行情感分类时，往往会因对上下文之间的语义考虑不全面导致文本情感分析的误差出现。

2.基于机器学习

基于传统机器学习的情感分析方法是一种通过模型来预测结果的学习方法，需要利用大量有标注或无标注的语料，使用统计机器学习的算法抽取特征，最后再进行情感分析来输出结果。该情感分析方法又可细分为三类：有监督、半监督和无监督。使用基于传统机器学习的情感分析方法主要在于情感特征的提取以及分类器的组合选择，不同分类器的组合选择会对情感分析的结果存在一定的影响。然而此方法在对文本内容进行情感分析时，会因存在忽略上下文语义、不能充分利用上下文文本语境信息而导致其分类准确性的下降。

3.基于深度学习

基于深度学习的情感分析方法主要依赖于神经网络的应用，使用预训练模型进行情感分析，选择合适的神经网络架构，再对大规模的文本数据进行训练，并对预测训练的模型进行微调和和验证从而实现更好的情感分析效果。相较于另两种分析方法，基于深度学习的方法通常能够实现更高的准确性，特别是在

处理大规模数据集和复杂的文本特征时尤为有效。因此本文情感分类主要基于深度学习技术，通过融合 CNN 神经网络模型与 Bi-LSTM 双向长短期记忆神经网络，并于模型中引入自注意力的方式，以得到更为准确的情感分类模型，此处模型原理于下文作详细介绍。

（三）爬虫技术介绍

网络爬虫（Web Crawler）是一种自动获取网页内容的程序，爬虫技术在数据分析、监控网站内容变化等多个领域发挥着重要作用。其基本工作原理是：首先选定一组起始的 URL 作为爬取的入口，通过 HTTP 或其他协议访问这些网页，解析网页内容，提取出新的 URL，然后再访问和解析这些 URL 指向的网页，如此循环，直到满足停止条件。

本文在进行情感分类研究的过程中，选取主流视频社交媒体 B 站作为数据源，通过 Python 第三方库 Bilibili-api 爬取用户对视频的评论，用于构建情感分析的数据集。

（四）深度学习相关技术

本文中主要基于深度学习技术对情感文本进行倾向性分析，受于篇幅限制，本节中仅详细介绍在下文中使用的深度学习网络。

1.卷积神经网络

卷积神经网络（Convolutional Neural Networks, CNN）是一类包含卷积计算的前馈神经网络（Feedforward Neural Networks），主要分为输入层、卷积层、池化层、全连接层和输出层五个部分，其主要模型示意如下图 2 所示：

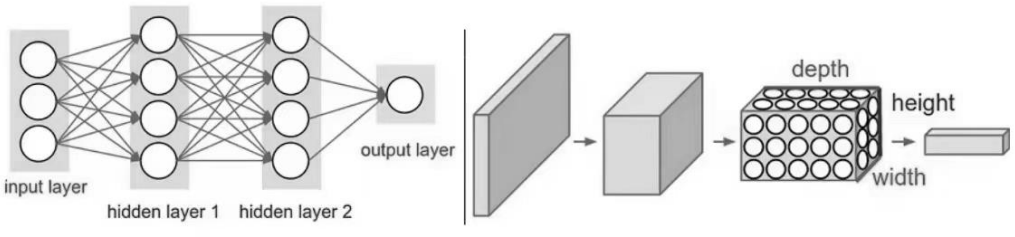

图 3 卷积神经网络 CNN 模型示意图

其中，输入层的主要作用是对原始预料进行预处理操作，方便后续的特征处理。假设输入长度为 N 的句子，设 W 为期望得到的词向量维度，则经过输入层可得 $M \times N$ 的词向量矩阵，即为卷积层的输入。

卷积层作为 CNN 的核心部分，主要用于特征提取。卷积层由多个卷积核构成，利用卷积核（kernel）在输入数据上滑动并进行点乘操作，生成新的二维数据，可用公式表示为：

$$S(i,j) = (I * K)(i,j) = \sum_{m}\sum_{n} I(m,n)K(i-m,j-n) \qquad \text{公式(1)}$$

其中 I 是输入图像，K 是卷积核，S 是生成的特征图，$*$ 则表示卷积操作。因卷积操作后得到的特征图可能存在维度较大的问题，以致引入池化层来减少参数矩阵的维度，从而避免过拟合现象的产生。

池化层为提取最重要的特征，对已提取的特征进行进一步的提取。最常见的池化操作是最大池化和平均池化，最大池化操作通过局部元素选择最大值来提取特征，而平均池化则是通过计算一定区域内所有元素的平均值来代表该区域的特征。最大池化的操作公式为：

$$P(i,j) = \max_{(m,n)\in R} S(m,n) \qquad \text{公式(2)}$$

平均池化的公式如下：

$$P(i,j) = \frac{1}{|R|}\sum_{(m,n)\in R} S(m,n) \qquad \text{公式(3)}$$

其中 R 表示池化区域, $P(i,j)$ 表示池化层输出特征图中的元素, $\frac{1}{|R|}$ 为取平均值时的系数。最大池化方法可以更好的保留信息的纹理特征而平均池化方法利于保留背景信息。

全连接层的作用是将前面层提取的特征图整合成最终的输出，全连接层的输出可以通过矩阵乘法和加权和来表示：

$$O = Wx + b \qquad \text{公式(4)}$$

其中 x 为输入向量，W 是权重矩阵，b 是偏置向量，O 代表输出向量。后再利用 softmax 函数分类全连接层的特征向量。

CNN 虽可有效提取局部特征，但因无法捕获长序列依赖关系和上下文的语义联系，易对情感倾向性的分析产生误差，因此提出长短期记忆神经网络。

2.长短期记忆神经网络

长短期记忆神经网络（Long Short-Term Memory Networks,LSTM）是一种特

殊的循环神经网络（RNN），能够学习长期依赖关系，其由 Hochreiter 和 Schmidhuber 在 1997 年提出，其网络架构与 RNN 高度相似，但 LSTM 的记忆细胞是其独有的结构，其旨在解决传统 RNN 在处理长序列数据时面临的梯度消失或梯度爆炸问题。

LSTM 的核心在于其内部的记忆细胞 $cell$，它可以分割长期信息与短期信息，同时赋予循环网络对信息做出选择的能力。在记忆细胞当中，LSTM 设置了两个关键变量：主要负责记忆短期信息、尤其是当前时间步信息的隐藏状态 h，以及主要负责长期记忆的细胞状态 C。

这两个变量都会随着时间步进行迭代，如下图 3 所示。

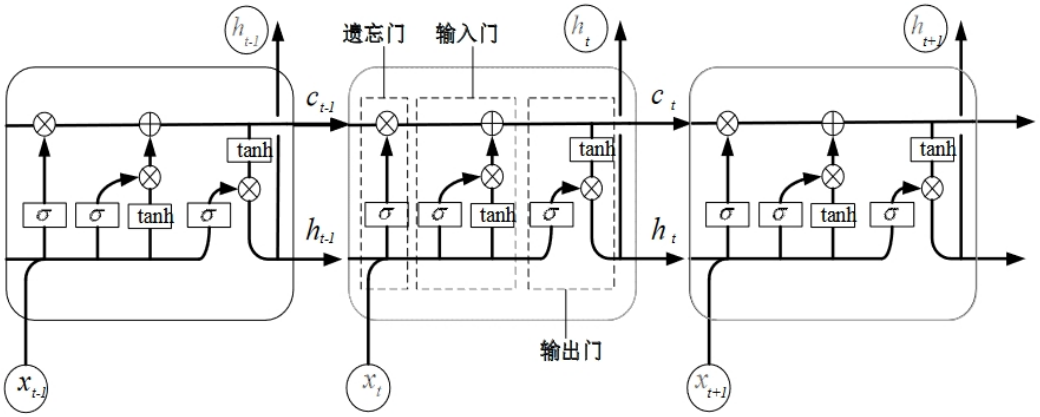

图 4 长短期记忆神经网络 LSTM 示意图

在迭代开始时 LSTM 会同时初始化 h_0 和 C_0；在任意时间步 t 上，记忆细胞会同时接受来自上一个时间步的 C_{t-1}、h_{t-1}以及当前时间步桑输入的新信息X_t三个变量，结合三者进行计算后，记忆细胞会输出当前时间步上的长期记忆C_t、h_t并将它们传递到下一个时间步上。同时，在每一个时间步上，h_t还会被传递到当前时间步的输出层，用于计算但前时间步的预测值 $\hat{y}_t$。

LSTM 每个记忆细胞都配有三个称为“门”（gate）的结构，这些门控制着信息的流入、保留和流出，分别为遗忘门(Forget Gate):决定哪些信息需要从单元状态中遗忘或丢弃。输入门(Input Gate):决定哪些新的信息要被添加到单元状态中。输出门(Output Gate):决定下一个隐藏状态（输出）应该包含哪些信息。在记忆细胞内部的计算流程图如下图 4 所示：

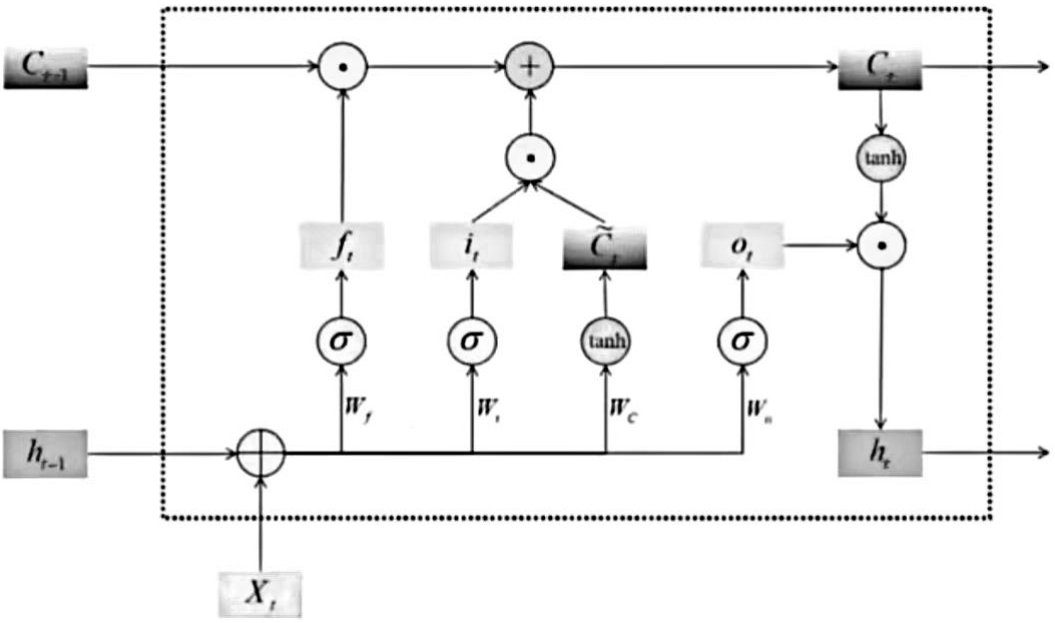

图 5 记忆细胞内部计算流程图

在第 t 时刻，不同的门的计算方式分别为：

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \quad \text{公式(5)}$$

$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right) \quad \text{公式(6)}$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad \text{公式(7)}$$

$$C_t = f_t \cdot C_{t-1} + i_t \cdot \tanh(W_c * [h_{t-1}, x_t] + b_c) \quad \text{公式(8)}$$

$$h_t = o_t \cdot \tanh(C_t) \quad \text{公式(9)}$$

其中 i_t，f_t，o_t，C_t分别代表第 t 时刻输入门、遗忘门、输出门和细胞状态的输出结果，其中x_t指第 t 时刻的输入，h_t为第 t 时刻区块中隐藏层中的向量，σ 表示 sigmoid 激活函数，其作用主要是选择性地过滤信息：输出值越接近 0 时趋于舍弃信息，而输出值越接近 1 则时则趋于保留信息，W 和 b 分别对应各个门中学习到的权重参数和偏置量。

LSTM 可以通过门控制机制较好解决 RNN 难以捕捉长距离依赖的问题，具有较好的记忆能力，但是因其传统的单向处理机制，导致其对上下文信息、文本的理解造成一定欠缺，因此引入双向长短期记忆神经网络（Bi-LSTM）。

3.双向长短期记忆神经网络

双向长短期记忆网络（Bi-LSTM）是一种特殊类型的循环神经网络（RNN）

架构，它能够处理序列数据并保持长期记忆，同时考虑过去与未来的信息，使得模型能够更好地捕捉序列数据中的上下文关系。

一个完整的 Bi-LSTM 网络包含输入层、前向 LSTM 层、反向 LSTM 层和输出层，其网络的一般结构如下图 5 所示：

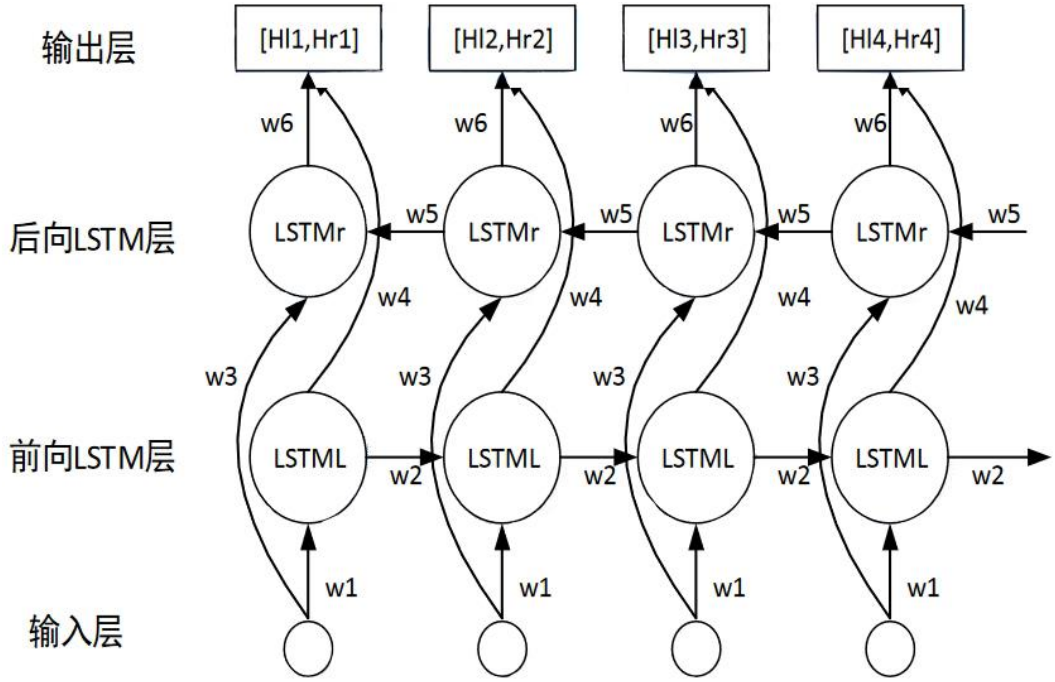

图 6 双向长短期记忆神经网络模型示意图

BiLSTM 模型中两个方向的 LSTM 都与之前介绍的结构相同，唯一不同的是在任一时刻 t，其隐状态 ht都是由正反两个方向的输出结果和此刻的输入信息共同决定的。用公式表达为：

$$(\widehat{h_t}) = LSTM_L\left(x_t, \widehat{h_{t-1}}\right) \quad \text{公式(10)}$$

$$(\widehat{h_t}) = LSTM_R\left(x_t, \widehat{h_{t-1}}\right) \quad \text{公式(11)}$$

$$h_t = w_t\overrightarrow{(h_t)} + v_t\overleftarrow{(h_t)} + b_t \quad \text{公式(12)}$$

上式中，第 t 时刻的输出为h_t，$\overrightarrow{h_t}$和$\overleftarrow{h_t}$分别代表 Bi-LSTM 中正向和反向的隐状态输出，LSTM结构的正向和反向分别为$LSTM_L$和$LSTM_R$，w_t、v_t分别代表正向和反向的权重系数，对应的偏置量为 b_t。

在处理长序列时，Bi-LSTM 可能会遇到性能下降的相关问题，因其记忆细胞容量有限，平等地处理整个序列上的信息。因此引出双向长短期记忆神经网络引入注意力机制（Bi-LSTM&Attention），加入的注意力机制使得模型得以更

加有效地聚焦于输入序列中的关键信息，提供了一种动态的信息重要性评估方式，提高了处理长序列和复杂依赖性的性能，解决了一部分 Bi-LSTM 的局限性。

4.注意力机制

注意力（Attention）可以进一步增强模型对重要信息的关注能力，在处理序列数据如文本、语音等方面表现出更优的性能。这种结合的模型不仅能捕捉时间序列的前后关系，而且能通过注意力机制聚焦于序列中最重要的部分，从而提高模型整体的解析和预测准确性。

注意力机制的本质是关注重点信息，忽略次要信息，在自然语言处理领域，注意力机制能够评估文本中各个单词与标签之间的相关性强度。通过分析查询向量（Query）和键（Key）之间的相似性，确定每个值（Value）的权重，从而计算出最终的注意力得分，其主要机制所下图 6 所示，主要由三个步骤计算所得：

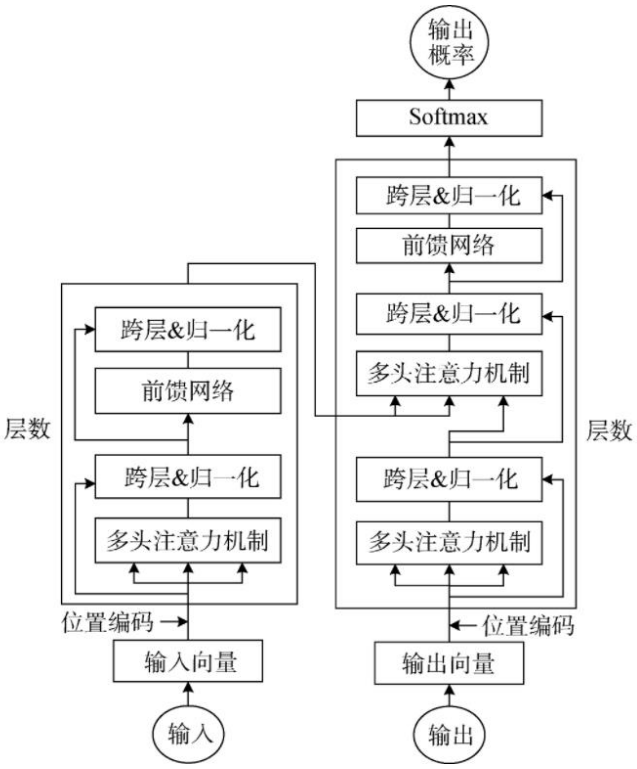

图 7 注意力机制示意图

①计算 Query 和 Key 的相似度，用e_{ij}表示：

$$e_{ij} = Query_i Key_j^T \qquad \text{公式(13)}$$

②利用 Softmax 函数进行归一化操作,使得 e_{ij} 最终以权重值的形式表示,数值范围为 0~1,计算如下：

$$cij = \frac{\exp(e_{ij})}{\sum_{k=1}^{N} \exp(e_{ik})} \quad \text{公式(14)}$$

③乘以矩阵 V ,得到最后的注意力得分 d_i：

$$d_i = \sum_{j=1}^{N} c_{ij} V_j \quad \text{公式(15)}$$

三、数据获取与处理

（一）数据来源及获取

近年来哔哩哔哩（Bilibili，简称 B 站），因其独特的社交属性及多样化的内容迅速成为年轻人乃至各个年龄段中愈发受欢迎的视频社交媒体，平台内容从起始的二次元内容为主，逐渐扩展到生活方式、科技等多个领域，吸引了广泛了用户群体。

由于该平台允许用户在视频下发表较长篇幅的评论，长篇幅的视频评价能够反映出更多信息，以及能够被更为精准的识别用户的情感态度，因此本文利用平台检索功能，以“小米 su7”“小米汽车”二者为关键词，按最多播放排序，爬取截止 5 月 8 日前，所有与研究主题直接相关且具有较大热度的视频评论，以广泛捕获用户对小米 su7 的情感倾向性；此处，由于 B 站的平台特性与用户习惯，各评论下的子评论较难准确反映用户情感特性，且识别不易，因此本文仅对各视频下的主评论进行爬取。

本次爬取的数据除用户评论外，同时对用户昵称、评论时间等基本信息进行获取，以便于后续数据清洗与舆情阶段划分处理。

（二）数据预处理

1.数据清洗

由于爬虫所获取的数据参杂了大量干扰信号，包括重复信息、广告、@某某，因此在进行正式分析前期，必须将这些信号全部剔除，以达到数据分析结论的准确性。

故本文首先对爬虫数据进行文本去重，以减轻数据中的信息冗余；同时利用 Excel 的“精准查找”对与主题无关的广告数据及@某某等无效数据进行人工剔除；此外，由于评论中包含大量表情，该类表情的存在一定程度上能够反映用户的情感倾向，因此对其进行直接剔除并不合适，本文将各类表情全部转化为能够被识别的英文字符，以处理表情对情感倾向的影响问题。清洗后的数据从 2024 月 1 月 7 日始至 2024 月 5 月 8 日止，共计评论数据 63739 条。

表 1 清洗后评论数据示例

序号	处理后正文
1	平板放车上，怕丢[doge]
2	小米一直上位，华为下位。小米性价比路线，华为营销奢侈品路线。两者不相上下
3	这次 su7 抄的真挺好，不像当年米 8 一样不伦不类，可惜穷哥们还是买不起，等混上去了高低买一台
……	……

2.文本分词与停用词去除

本文于分词阶段，选用 python 第三方库 jieba 进行文本分词处理，其作为目前国内著名的分词工具，被广泛应用于中文文本分析，其共具有 3 种模型：适合文本分析的精确模式、速度快但是不能解决歧义的全模式和适合搜索引擎分词的搜索引擎模式。本文需要对情感进行准确的倾向分类，因此选择了 jieba 分词的准确模式，将句子最精确的切开。将清洗后的的 B 站评论数据作为输入，分词结果作为输出形成新的一列，记为“tokens”。

经过上述文本分词后，数据中仍存在大量频率较高，却没用实际意义的词语，本文中主要为“小米”“su7”“汽车”等对文章主题的确定没有帮助的词汇；以及各句子中的语气助词，如“的”“呀”等，本文中将以上两种类型的词汇全部进行剔除，得到最终分词后的文本数据，用于下文调用。

表 2 分词后数据示例

序号	Tokens
1	厦门 新车 行驶 39 公里 出现 故障 小米 服务中心 表示 无法 维修 更换 全额 退款 故障 知道
2	雷总 加油 早日 追 特斯拉 贷款 买 20 万 丢人 没错 朋友
3	外观 吸引力 内在 更 重要 企业 长久 关键 可能 一直 抄袭 迟早
……	……

四、阶段性舆情聚焦分析

（一）小米 su7 舆情阶段性划分

为更为准确的把握本次小米 su7 发售的网络舆情影响情况，本文基于百度搜索指数趋势图，对本次小米 su7 的舆情事件进行阶段性划分。

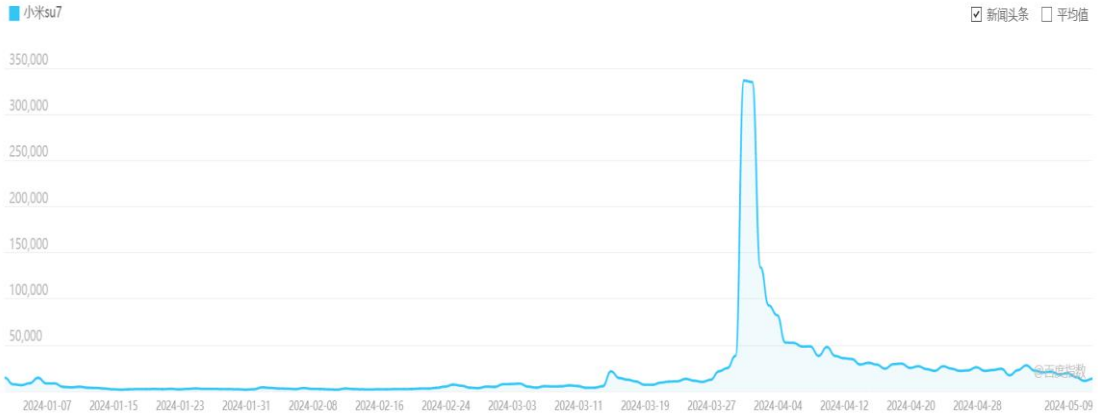

图 8 小米 su7 百度搜索指数趋势图

由上趋势图可知，小米 su7 的舆情爆发期较短，主要集中在 3.28 日至 2.29 日期间，该时间段处于小米 su7 的上市发布会召开期间，舆情出现爆发式增长。而此前小米 su7 的舆论的指数相对平缓，并无激烈波动，且自舆情爆发期后，小米 su7 的舆情迎来迅速衰退，并保持相对稳定。由此，本文将小米 su7 的上市发布会召开作为节点，将舆情阶段进行切分，得到如下三阶段：2024 年 1 月 7 日至 2024 年 3 月 28 日晚七点，认为是舆情预热期；2024 年 3 月 28 日晚七点至 3 月 29 日认为是舆情爆发期，3 月 30 日至 5 月 8 日，认为是舆情衰退期。基于此，本文以舆情阶段作为分析基础，以更为准确地把握用户的舆情倾向度。

（二）基于词云分析的阶段性聚焦点探索

本文首先对各阶段进行高频词统计，用以词云图的绘制，以直观的形式，从整体上把握用户对小米 su7 的关注侧重点。

1.舆情预热期

舆情预热阶段，小米 su7 的“价格”是用户的主要关注点，考虑到此前小米凭借平价的产品在科技行业内享有较高的知名度，因此用户对小米 su7 的价格尤为重视，“20”“30”等数字词语与“性价比”等词汇，正是用户对小米 su7 上市价格的预测与期望；同时，作为手机起家的小米，“手机”也被用户多次提到，用户对小米在汽车圈内的定位也较为关注；此外，“比亚迪”“电车”也多

次被提及，小米 su7 作为电车行业的新人，势必受到与扎根该行业较久的汽车品牌间的比较，这也对小米汽车造成了巨大挑战。

图 9 舆情预热期词云图

2.舆情爆发期

舆情爆发阶段，“手机”转而成为用户提及最多的词汇，此处除了小米作为传统的手机厂商引起了较大的关注外，发布会上所提到的例如“小米 su7 提供行业内最好的车内手机支架”等营销宣传也有一定关联；同时“定金”“大定”此类与订购小米汽车相关的词语也在该阶段被大量提及；此外，“价格”等相关词语同样在该阶段备受关注，但值得注意的是“营销”等负面关键词一定程度上也被用户所提及。

图 10 舆情预热期词云图

3.舆情衰退期

舆情衰退阶段，“问题”是在用户群体中提及最多的词汇，考虑是小米 su7 在发布后，出现了大量关于小米 su7 的车辆问题，以及在大量对小米 su7 的讨论中，显现出用户对小米 su7 的消费信任度较低问题；但与此同时，“喜欢”这类较为积极的词语也呈现较高占比，同样能够说明小米 su7 是抓住了部分用户群体的汽车偏好的；此外，由于发布会中雷军将小米 su7 对标汽车品牌“保时捷”，因此与“保时捷”品牌相关的词语也被大量提到。

图 11 舆情预热期词云图

（三）基于 LDA 主题分析的阶段性话题归纳

1.模型选取与评估指标确立

基于上文词云分析，本文对消费者在各阶段的关注点做了整体把握，但由于词云属于粗粒性分析，故本节基于 LDA 主题分析对各阶段进行话题归纳，以更为细致地掌握各阶段消费者的具体关注情况，本文将使用 gensim 包构建 LDA 模型，将清洗完毕的文本分词数据作为参数传入，作为模型的输入。

对于主题模型，评估其效果的优劣非常关键，尤其是在没有标记数据的情况下。故本文将基于两种重要的主题模型评估方法：主题困惑度（Perplexity）和主题一致性（Topic Coherence），以综合评估 LDA 模型的性能，以确定最优主题数。

主题困惑度是评价 LDA 主题模型的一个常用指标，主要用来衡量模型对新

文档的预测能力，其困惑度越低，意味着模型的预测性能越好。困惑度的定义是模型预测样本的概率的几何平均的倒数，可通过如下数学表达式计算：

$$\text{Perplexity}(D) = \exp\left(-\frac{\sum_{d=1}^{M}\sum_{n=1}^{N_d}\log p\,(w_{dn})}{\sum_{d=1}^{M}N_d}\right) \quad \text{公式(16)}$$

其中，w_{dn} 是文档 d 中的第 n 个单词，M 是文档的总数，N_d 是文档 d 中的单词数，$p(w_{dn})$ 是模型关于单词 w_{dn} 的预测概率。

主题一致性评价的是一个主题内的词语是否具有高度相关性，通过评估词语之间的统计相关度来实现，主题一致性得分高表明主题内的词语紧密相关，反映了主题的质量。本文基于 gensim 库中的 c_v 一致性得分进行计算，该方法综合考虑了主题词汇的统计共现关系与语义相似。

其利用滑动窗口方法，计算每一词对在整个语料中共同出现的概率 $P(v,u)$，以及各自出现的概率 $P(v)$ 和 $P(u)$，使用归一化点互信息（NPMI）作为确认度量，公式如下：

$$\text{NPMI}(v,u) = \frac{\log\frac{P(v,u)}{P(v)P(u)}}{-\log P\,(v,u)} \quad \text{公式(17)}$$

这里，$P(v,u)$ 是词 v 和 u 同时出现的概率，而 $P(v)$ 和 $P(u)$ 是单独出现的概率。

最终的c_v一致性得分为所有词对 NPMI 值的算术平均值，表达式为：

$$C_V = \frac{1}{N}\sum_{v\in V}\sum_{u\in U}\text{NPMI}(v,u) \quad \text{公式(18)}$$

其中，N 是所有考虑的词对数量。

2.舆情预热期主题特征

首先计算并绘制出舆情预热期的 20 个模型的困惑度与一致性曲线，以确定最优舆情话题数，如下图 11 所示：

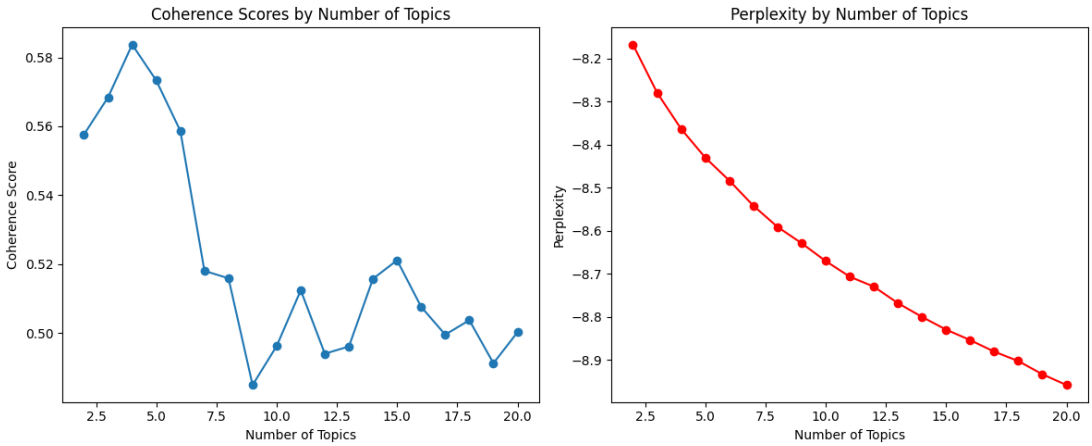

图 12 舆情预热期的困惑度与一致性曲线

其中，由于本文采用log_perplexity()函数计算困惑度，其得到的困惑度为实际困惑度的对数表达式，得到的困惑度并非实际困惑度，需按如下公式转换才能得到真实的困惑度：

$$perplexity = 2^{-\log\_perplexity()} \quad \text{公式(19)}$$

由该公式可知，原始困惑度数据值越高，实际困惑度则越低，模型效果越优，本阶段中，基于真实困惑度水平与一致性曲线，并结合气泡图使得各主题间尽量无重合，可综合判断得其最佳主题数为6，由此得到最终主题类别及各主题词分布，取分布最高的 5 个主题，各主题中取 6 个关键词进行分析，分析结果如下表 1 所示。

表 3 舆情预热期 LDA 主题分析结果

主题序号	主题类别	类别占比	关键词
Topic1	品牌影响	46.3%	价格、手机、比亚迪、电池、性价比
Topic2	营销手段	20.6%	价格、营销、电车、成本、电池
Topic3	品牌价值	11%	市场、电车、价格、品牌、手机
Topic4	品牌技术	9.4%	便宜、比亚迪、价格、产品、技术
Topic5	市场战略	6.8%	华为、技术、产品、价格、发市场

由此可知，在舆情预热期，消费者主要聚焦在小米品牌自身的讨论上，受其品牌战略的影响，大多数消费者尤其关注小米汽车在价格上的表现，并且多数消费者对小米品牌的印象，仍停留在传统的手机厂商上，要改变消费者对小米汽车的印象依然任重道远；同时小米的营销手段也是消费者关注的一大侧重点，小米凭借其出色的营销手段，吸引了大批消费者的关注；此外，在舆情预

热期，出现了许多小米品牌与其余汽车品牌之间的对比，消费者在不同车企之间的衡量，也对小米汽车后续的上市发售造成了不小压力。

3.舆情爆发期主题特征

按相同标准同样计算并绘制舆情爆发期的 20 个模型的困惑度与一致性曲线，如下图 12 所示。

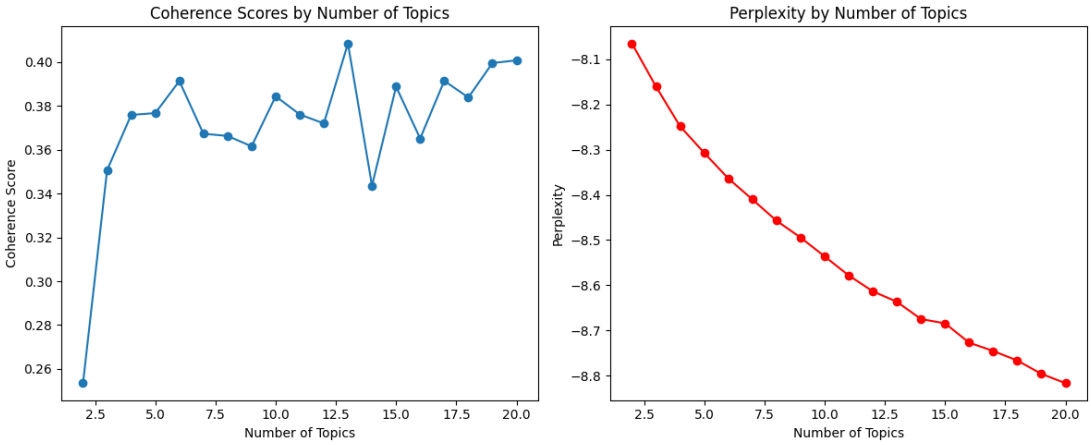

图 13 舆情预热期 LDA 主题分析结果

由上图进行综合判断，并结合气泡图使得各主题间无重合可知，舆情爆发期的最优话题数为 7，同样取分布最高的 5 个主题，各主题中取 6 个关键词进行分析基于此话题数得到如下 LDA 主题分析结果，如下表 2 所示。

表 4 舆情预热期 LDA 主题分析结果

主题序号	主题类别	类别占比	关键词
Topic1	发售上市	37.7%	发布会、价格、支持、电车、滑稽
Topic2	汽车预定	21.7%	定金、手机、价格、米粉、意思
Topic3	设计对比	13.4%	保时捷、小爱、设计、买不起、电池
Topic4	产品特色	8.6%	影响、便宜、性价比、手机、智能
Topic5	售后服务	8.5%	质保、电池、座椅、空间、电机

由此可知，舆情爆发期的话题主要聚焦于发布会本身，讨论范围涵盖了汽车发售上市、汽车预定、设计对比、产品特色以及售后服务等多方面的内容。该阶段的主要讨论话题是小米汽车的发售上市，语料分布占比 37.7%，消费者最为关注小米在发布会上提及的关于小米 su7 的各项数据；同时，关于小米汽车预定的话题也被多数消费者关注，如何支付定金，预定小米汽车的词汇词频较高；再者，小米汽车的设计与产品特色也备受关注，“小爱”等具有小米特

色的功能被讨论较多；此外，小米作为新能源行业的“新人”，其售后服务也受到广泛关注，小米是否能提供不弱于老牌新能源汽车品牌的售后服务，其供应链能否满足消费者需求，也对小米汽车造成了一定的挑战。

4.舆情衰退期主题特征

同样首先通过困惑度与一致性曲线确定最优化数，如下图 13 所示。

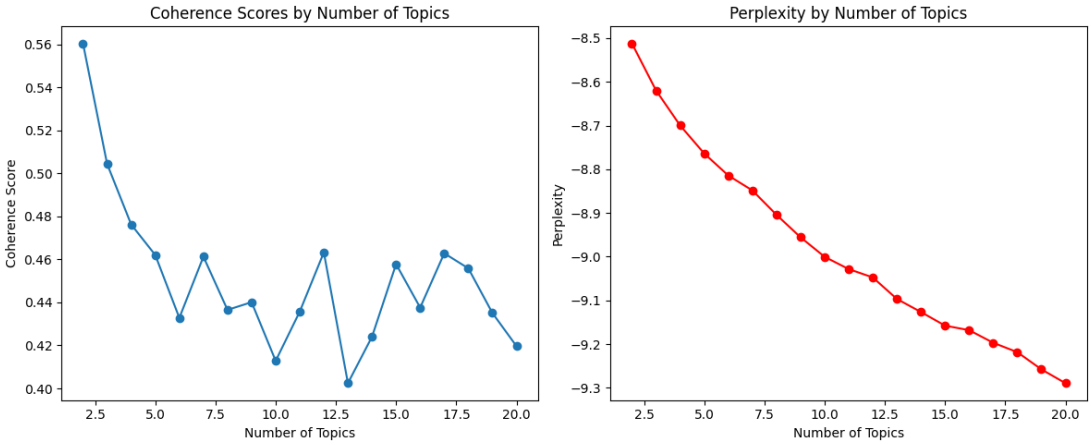

图 14 舆情衰退期的困惑度与一致性曲线

由上图综合判断，并结合气泡图使得各主题间无重合，以此确定最优话题数为 5 个，取各主题中的 6 个关键词进行分析基于此话题数得到如下 LDA 主题分析结果，如下表 3 所示。

表 5 舆情衰退期 LDA 主题分析结果

主题序号	主题类别	类别占比	关键词
Topic1	退订风波	48.4%	米粉、问题、价格、定金、营销
Topic2	消费体验	21%	模式、保时捷、智能、手机、体验
Topic3	外观设计	12.7%	保时捷、设计、空间、外观、后排
Topic4	性能配置	10.8%	续航、电池、性能、智能、差距
Topic5	销售状况	7.1%	销售、质保、国产、车型、座椅

由上可知，主题退订风波在整体语料分布中占比较高，主要是由于小米在发布会后瞬间迎来了大量的预定潮，但由于电车行业的一些规则所限，预定的汽车原则上不允许退订，因此给小米汽车造成了大量的舆论风波，但这也从侧面反映出，小米汽车的发布受到了大量消费者的青睐；而整体上其余舆情衰退期的话题主要聚焦于消费体验上，主要是由于在发布会过后，大量的线下测评视频开始出现，消费者开始更为关注小米汽车的试驾体验。其中，小米汽车带

来的智能化体验、多样化的外观设计、优秀的性能配置成为多数话题的主要讨论聚焦点；此外，在舆情爆发期中备受消费者关注的销售担忧，也在该阶段受到大量讨论。

（四）本章小结

本章介绍了基于词云分析与 LDA 主题特征建模的“小米汽车 su7”事件的舆情聚焦点。首先通过对舆情事件进行阶段性划分，将各舆情阶段作为本章聚焦分析的基础；再者，基于此对各舆情阶段进行词云分析，以整体把握消费者在各阶段的舆论聚焦点；最后基于 LDA 主题建模，对各舆情阶段进行更为细致的话题归纳，并对各阶段的舆情话题进行概括分析。

五、阶段性情绪动态演化分析

本文在第四章中基于舆情事件的阶段性划分，对各阶段舆情进行了聚焦点分析，但限于仅通过主题特征分析，难以知晓消费者的情感倾向性态度。因此本章中基于一种引入注意力机制的融合 CNN 与 Bi-LSTM 神经网络的模型，对用户视频评论进行舆情情感分类，并基于分类后的用户情感特征，结合舆情阶段与主题特征建模，对舆情事件的动态情绪演化进行分析，以探究消费者对小米汽车的情感倾向性态度。

（一）基于 CNN-Bi-LSTM-ATT 模型的舆情情感分类

1.CNN-Bi-LSTM-ATT 模型搭建

本章利用一种将 CNN 与 Bi-LSTM 相结合并融入注意力机制的模型，以提高对情感分类任务上的准确度，构建出的 CNN-Bi-LSTM-ATT 模型结构如下图 14 所示：

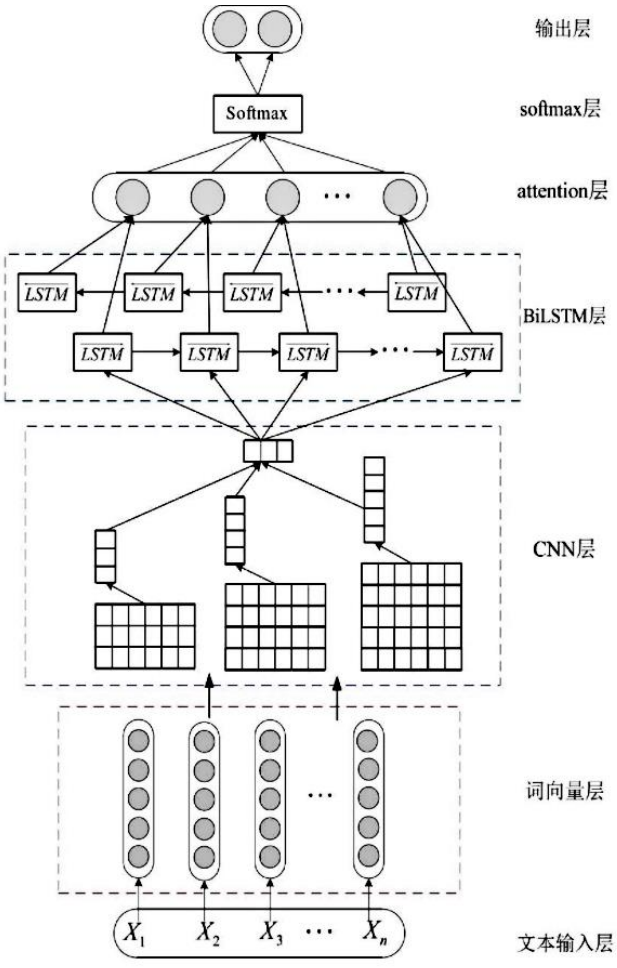

图 15 CNN-Bi-LSTM-ATT 模型示意图

该模型主要包含七层结构，首先通过文本输入层，对文本语料集中的数据进行预处理操作，再通过词向量层对预处理后的文本数据进行词向量化处理，处理后的数据通过 CNN 对局部特征进行提取，利用 Bi-LSTM 捕获上下文的序列特征，将两者相融合，使得对文本特征的获取更为全面，再引入注意力机制聚焦序列中最重要的部分，进一步提高模型整体的预测准确性，最后通过 softmax 层进行情感分类决策。

本文在利用 CNN 进行局部特征的提取时，因 CNN 中包含多个按顺序叠加的隐藏层所组成的特征提取器，能够在不同的层级中提取出各种局部信息，减少显示的特征提取，降低模型的复杂度，在一定程度上可防止过拟合现象的产生。在文本处理中，CNN 通过使用卷积核有效地提取文本的局部语义特征，从而识别出文本的情感属性并进行情感分类，该模型中 CNN 网络的原理介绍如下，其余层与本文第二章第四节中介绍的原理相近，不作过多阐述。

将词向量层的输出作为输入，假设 n 为次数，每个词对应的词向量为 x_i，卷积核为 w，卷积核大小为 g，则 i 到 $i+g-1$ 的词向量为 $x_{i:i+g-1}$， b为偏向量，可得特征矩阵$W=[W_1,W_2,W_3,\cdots,W_{n-g+1}]$：

$$W_i = f\left(w \times x_{i:i+g-1} + b\right) \qquad \text{公式(20)}$$

采用最大池化技术，通过对经卷积操作所得的局部特征矩阵进行下采样，以提取各局部区域的最大值，从而获得最显著的特征。

$$x = \max(W) \qquad \text{公式(21)}$$

特征最后通过全连接层进行整合，将池化后的分散特征序列x_i合并成一个新的连续向量 Y：

$$Y = (x_1 + x_2 + x_3 \cdots ,\ x_n) \qquad \text{公式(22)}$$

向量 Y 包含较高层次的数据，将作为双向长短时记忆网络层的输入。

2.小米汽车 su7 数据集的构建

本文中通过爬虫技术所爬取的评论数据，其自身并无情感倾向性标注，因此利用上文所述的 CNN-Bi-LSTM-ATT 模型进行情感预测，但鉴于目前并无小米汽车 su7 领域的特定数据集供使用，且传统通用数据集对其训练效果，会由于领域的特定性产生偏差，因此本文选择通过构建一个小米汽车领域的特定数

据集以用于更多该领域数据集的情感倾向性预测。

目前对于特定领域数据集的构建，大多选择以人工标注的方式进行，但限于人力所限，且对于大量非结构化数据文本的标注，人工识别在标注后期极易导致情感分类的偏差，因此本文选择采用伪标签法，以训练一个在小米汽车领域的特定数据集。

①通用数据集的获取

由于传统的通用数据集难以满足本研究对准确情感分类的需要，因此本文通过人工标注的方式，得到 8000 条准确的情感分类数据集，并通过这小部分准确分类的数据集基于伪标签法，在特定领域情感样本较少的情况下，以训练得到一个更大的小米汽车领域的特定数据集。

②伪标签法

伪标签法（Pseudo-labeling）是一种半监督学习方法，通常用于利用未标记数据来改善模型的学习性能，其特别适用于在有限的标记数据下的训练任务，以此可获取大量未标记数据。

其基本思想是利用已有的标记数据训练一个初步模型，然后使用这个模型对未标记数据进行预测，生成伪标签。随后，将这些带有伪标签的数据再次用于训练模型，以此来拓展训练样本的规模和多样性。整个过程会迭代多轮，每一轮都希望通过增加更多的伪标签数据来提升模型的性能。

③小米汽车 su7 领域特定数据集构建流程

本小节首先将使用人工标注得到的 8000 条准确分类的情感数据集训练一个性能最优的情感分类器，该情感分类器选用被广泛使用的 Bi-LSTM-ATT 模型，得到训练后的模型后，对爬取的用户评论数据进行预测，通过阈值筛选，将筛选后分类准确的评论与通用数据集合并，且再次训练模型，并重新进行预测，重复这个过程，直到得到一个情感分类较为准确的小米汽车 su7 数据集停止。构建实现流程如下图 15 所示。

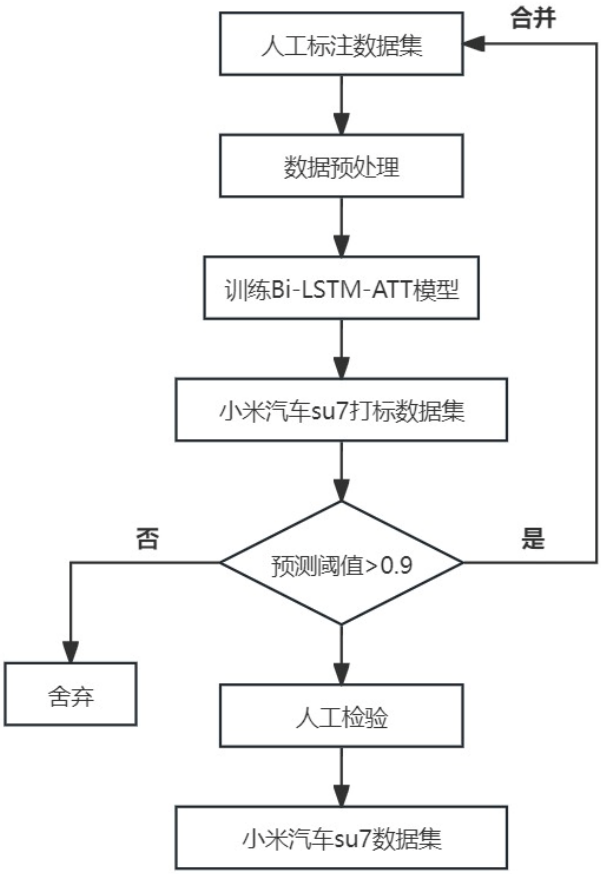

图 16 小米汽车 su7 领域评论数据集构建流程图

具体流程如下：

（1）通过人工标注的方式得到 8000 条准确分类的情感分类数据集，尽可能保证数据集中情感倾向均衡，以达到较好的分类效果，并以 3：1：1 的方式划分训练集、测试集以及验证集，得到训练集共 4800 条，测试集与验证集均为 1600 条；

（2）对得到的 4800 条训练集，按上文所述的方式进行数据预处理，以用于对 Bi-LSTM-ATT 模型的训练；

（3）搭建 Bi-LSTM-ATT 模型，对处理后的数据进行训练，得到初始的情感分类器。

（4）从待分类的评论数据中选取 24000 条，用于预测，以构建预测数据集，用训练完毕的 Bi-LSTM-ATT 模型对这 24000 条数据进行情感预测，得到打标数据；

（5）设定筛选阈值为 0.9，保留超过该阈值的评论打标数据，并将这部分数据集与通用数据集合并；

（6）使用（5）中合并后的数据集，再次对 Bi-LSTM-ATT 模型进行训练，并重新对待打标数据进行预测，且进行阈值筛选；

（7）重复上述步骤，直至所有打标数据均通过阈值筛选，或迭代次数超过 10 次，则流程停止。

（8）对最后一次训练得到的 Bi-LSTM-ATT 模型性能进行评估，该分类器的 Accuracy：98.5%、Recall：98.56%、F1 值：98.58%，分类效果较好，能够准确对特定领域进行情感分类；

（9）对最终得到的小米汽车 su7 特定领域的数据集，进行人工核验，确保数据集能够正确表达用户的情感倾向。

（二）阶段性情绪演化分析

在上文基础上，我们将得到的小米汽车 su7 特定领域的数据集与人工标注的数据集进行合并，得到最终的情感分类数据集，将此数据集以 8：1：1 划分为训练集、测试集以及验证集，将训练集输入至 CNN-Bi-LSTM-ATT 模型中，利用该模型对剩余用户评论数据进行预测（包括伪标签法训练前期舍弃的数据），CNN-Bi-LSTM-ATT 模型 Accuracy， 95.23%、Recall：95.85%、F1 值：95.25%，分类性能较好，至此剩余所有评论数据情感分类完毕，共 63739 条。

其中，积极情绪 16676 条，中性情绪 26090 条，消极情绪 20973 条。整体而言，三种情绪占比较为均衡，消费者中对小米 su7 更多持消极态度，对其批评比例相对较高，为更详细地把握消费者情绪特征，下文将基于上文的情感分类结果进行情绪演化分析，以动态地把握消费者情绪的变化特征。

计算各阶段不同情绪占该阶段的情绪占比，得到如下图 5-3 所示情绪演化折线图。

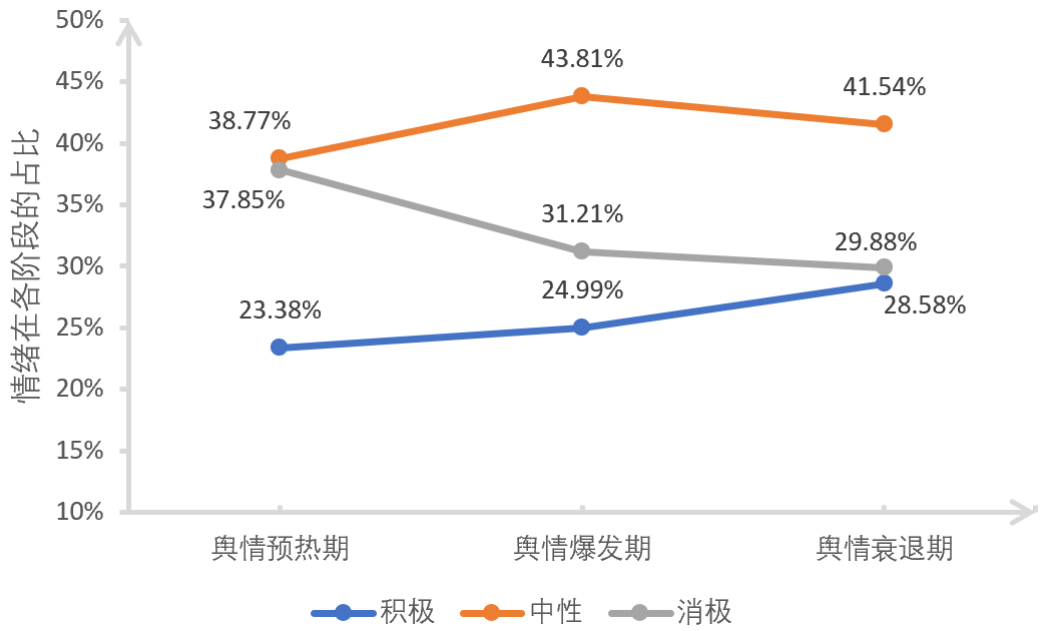

图 17 各阶段不同情绪占比演化图

就其积极情绪与消极情绪的动态演化而言，其积极情感随着舆情阶段的变化，呈现相对平稳的上升趋势，而消极情绪则呈现逐渐下降的趋势。这从一定程度上说明了，随着小米 su7 逐渐亮相，各项配置呈现在大众视野中，消费者对小米 su7 的情绪倾向从一开始的相对不看好，逐渐过渡到消费信心的逐渐增强。同时也需注意到，从整体看，舆情预热期到舆情衰退期，消费者对其的积极情绪一直低于消极情绪。但是，在最后的舆情衰退阶段，其积极情绪与消极情绪十分接近，也说明了在热度逐渐过去后，消费者对其的态度呈现分化对峙，对其的平价呈现两极分化。因此小米汽车 su7 仍然在消费者中仍然备受争议，小米品牌仍需继续谋求发展，以转变消费者对其的情绪倾向态度。

此外，其中性情感而言，其在舆情预热期表现最低，占比 38.77%，而在舆情爆发期达到顶峰，达 43.81%，舆情衰退期略有下降，但仍保持较高水平。整体而言，消费者在各个阶段中，持中性情感态度的最多，对小米汽车 su7 的情感倾向相对理性。这对于小米品牌而言，是一个十分重要的信号，小米汽车能否在接下来的产品中转变这部分消费者的情感态度，对于其未来发展，在消费者端有重要意义。

（三）阶段性情绪 LDA 主题特征演化

由上阶段性情绪演化图，本文对不同阶段下消费者对小米 su7 的情感倾向

性做了整体阐述，为进一步探究不同阶段下的情感倾向性特征演化趋势，本文基于 LDA 主题分析，对三个阶段下的各个情绪进行 LDA 主题特征探索，依据上文确定最优主题数的方式，得到共 9 个主题分析特征结果，该步骤与上文类似，不做过多展示。根据该结果绘制如下文所示桑基图，以反映情绪特征间的动态演化趋势。

1.积极情绪阶段性演化

就其积极情绪而言，在舆情预热期，主要聚焦于产品特性与技术创新，其积极情绪主要从这两个主题过渡到产品的外观设计、智能科技以及外观设计，并在舆情衰退期，侧重关注产品细节与消费体验。整体而言，消费者的积极情绪变化主要受小米汽车的产品特性影响，对产品细节的消费体验很大程度上会影响其积极情绪的表示。因此小米品牌需要重视对产品细节的打磨，给予消费者更佳的消费者体验，有助于其刺激消费需求。

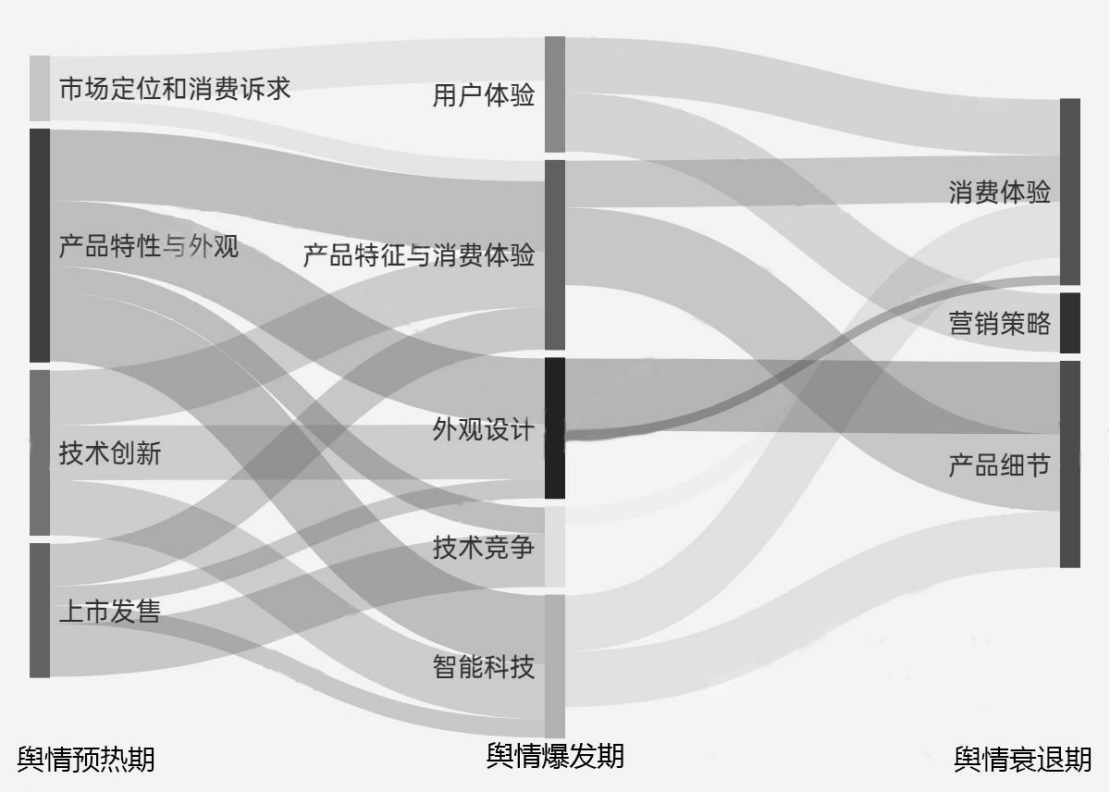

图 18 积极情绪的桑基图

2.中性情绪阶段性演化

就其中性情绪而言，舆情预热期的主要焦点其中在车企间的竞争与社交影响上，这一方面体现了消费者常将小米汽车与成熟的车企进行比较，同时也显

示出他们对于小米汽车对消费者在社交领域的影响持重视态度；而舆情爆发期，其聚焦点则重点转向于产品发布与售后服务，小米 su7 的上市发布会的召开引起了对产品本身的大量讨论度，同时小米作为车企领域的新进者，也引起了消费者对其售后服务是否完善的大量关注；在舆情衰退期，中性情绪的变化则过渡体现在小米汽车的电车性能与客户体验上，对小米 su7 的性能配置以及试驾体验更多表现出中性情绪。整体而言，消费者的中性情绪更多体现在对小米汽车的消费体验上，因此小米品牌应重视这部分中性情绪的消费者的关注侧重点，该类人群是小米汽车的重要潜在消费者。

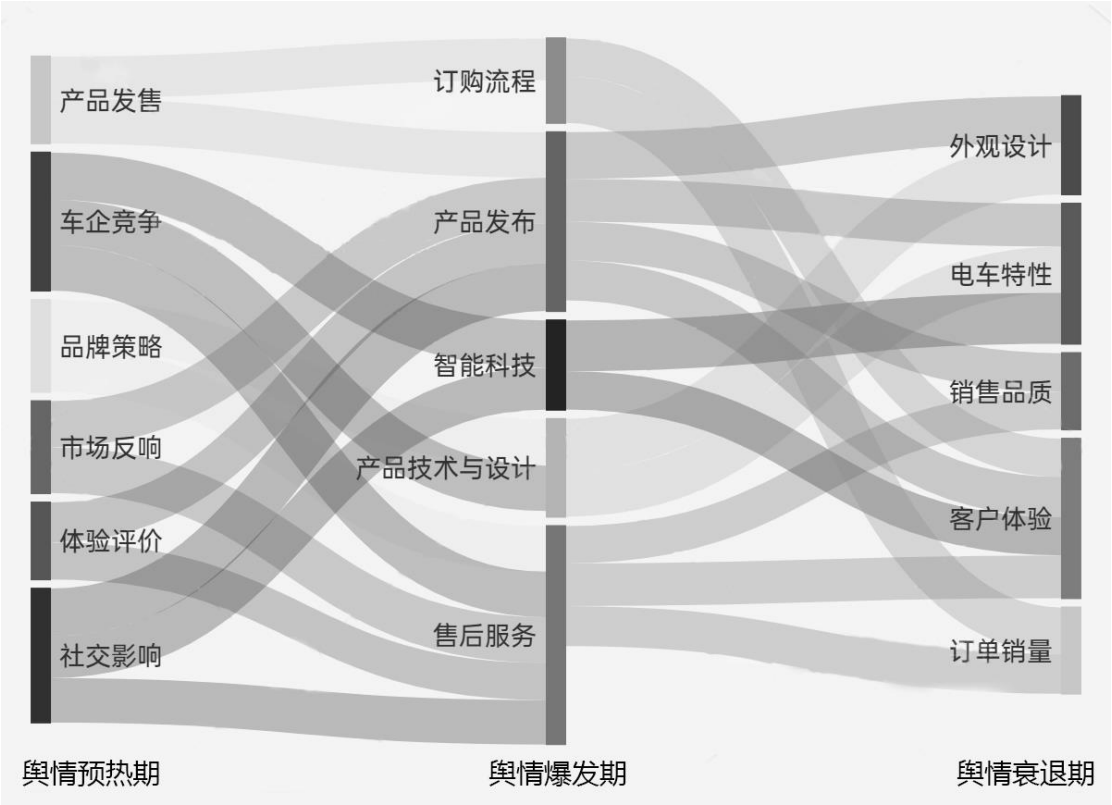

图 19 中性情绪的桑基图

3.消极情绪阶段性演化

就其消极情绪而言，在舆情预热期，消费者的负面情绪聚焦点较为分散，定价策略、供应链不成熟等多方面的因素，都使得消费者表现出消极情绪。小米品牌加大研发投入，在各个领域上减轻消费者的消费担忧。而到舆情爆发期，其消极情绪影响因素，主要转向对技术品质、消费体验以及对营销宣传的不满上。消费者对小米汽车在性能配置、车辆体验的质疑， 以及对过度批量营销的方案，成为了该阶段引起消费者负面情绪的主要聚焦因素。到了舆情衰退期，

消费者的负面情绪滋生主要受价格策略影响，考虑在发布会后，多方线下销售渠道开启，未能符合消费者预期的定价，使得成为该阶段预期的主要诟病。此外，外观设计与销售服务等与线下订购息息相关的影响因素，都一定程度上对消费者的负面情绪产生了影响。因此建议小米汽车加大研发投入，以成熟的汽车工艺反哺高性价比的价格策略，以减轻消费者反感，以此建立良性循环。

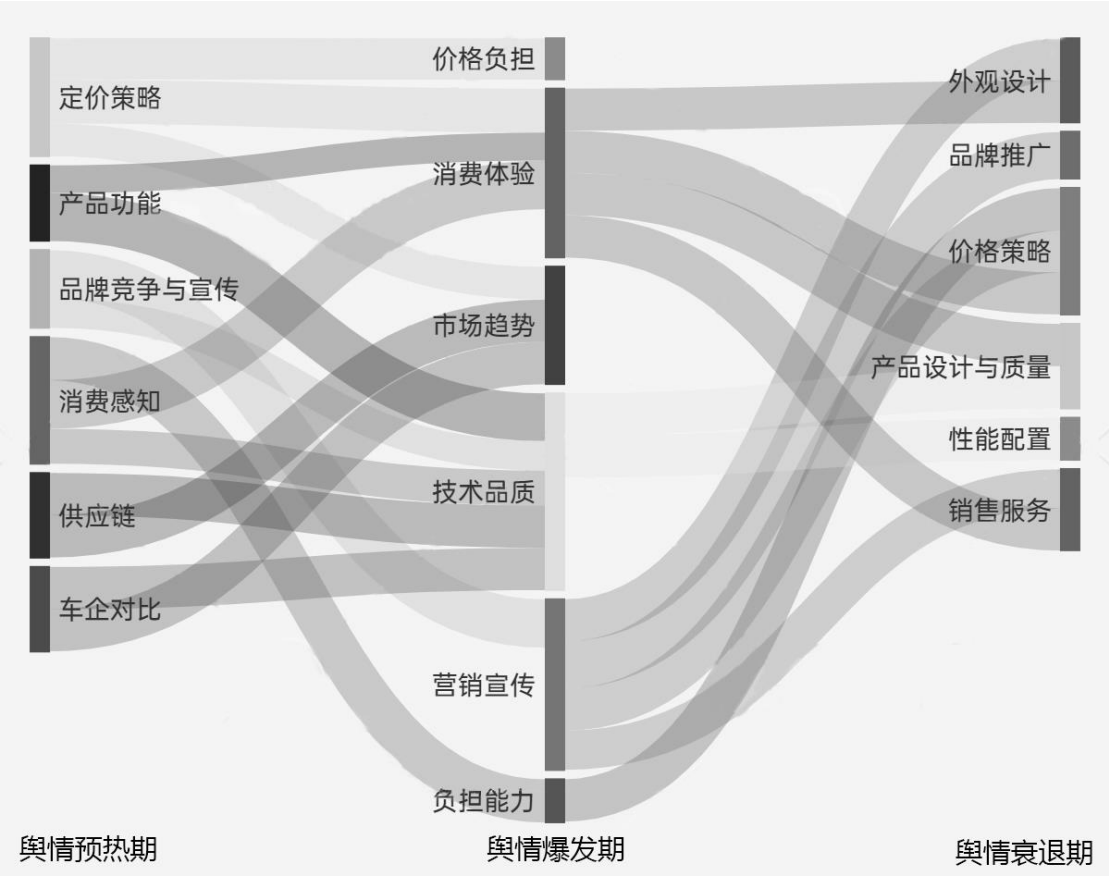

图 20 消极情绪的桑基图

（四）本章小结

本章通过伪标签法构建适用于本文特定领域的小米汽车数据集，并采用一种引入注意力机制且结合 CNN 特征提取的 Bi-LSTM 模型，对用户评论数据进行情感分类，基于舆情的阶段性划分，对消费者在各阶段下不同情绪的演化特征进行分析，探究消费者情感倾向转变的聚焦因素，以此对小米汽车的发展提供一定建议。

六、总结与不足

（一）工作总结

本文主要基于主题特征提取与深度学习方法，对小米 su7 的舆情中的消费者关注聚焦点以及情感倾向性进行探究，具体工作内容如下：

①利用爬虫技术对本文所用数据集进行爬取，对获取数据文本进行清洗、去重、分词等预处理工作，并对文本数据中难以处理的表情倾向，转换为可被识别的字符表示，使处理后的文本适用于 LDA 主题模型与神经网络的输入。

②基于百度指数的舆情演化趋势，将舆情切分为舆情预热期、爆发期以及衰退期三个阶段，下文工作皆基于不同阶段下的舆情内容进行探究。

③基于词云分析对不同阶段下的舆情进行整体探讨，并进行 LDA 主题建模，以更为细化地探究各阶段下的不同话题聚焦点。

④利用伪标签法构建小米特定领域的数据集，并基于此采用一种引入注意力机制且结合 CNN 特征提取的 Bi-LSTM 模型，对用户评论数据集进行情感分类，得到较为精准的情感分类数据集。以此情感分类数据集为基础，探究不同阶段下的消费者情绪演化，以洞察不同阶段下影响消费者情绪倾向的聚焦特征。

（二）研究结论

基于如上工作，本文得出如下结论：

①消费者对小米汽车的主要舆情聚焦点，在于对小米品牌新进车企行业的消费信任感缺失，该类担忧贯穿了小米汽车舆情的三个发展阶段，各个阶段中，消费者对其的担忧侧重点有所转变，但仍主要集中于定价策略、营销手段、性能配置以及外观设计四个方面。

②消费者在舆情中的情感态度，整体持中立，大多数消费者对小米汽车的态度评价相对理性，但消极情绪在三个阶段中所持的比例，均高于积极情绪，有相当部分消费者对小米汽车的发售持怀疑态度，其聚焦因素在不同阶段呈现多样性。但对于舆情的最后阶段，在小米汽车逐渐亮相于大众视野中后，积极情绪一路上升，直至与消极情绪的比例极为接近，证明小米汽车的发售得到了很多消费者的认可，也同样改变了许多消费者的刻板态度。但即使是舆情的最后阶段，其舆情仍然呈现出情绪的极端分化。

（三）对策建议

基于上述舆情探究，本文对小米品牌提出如下对策建议：

①小米汽车作为车企行业的新进者，其第一辆车的发售结果，虽没有得到舆论的广泛认可，但舆情的分化使得小米汽车应格外重视其下一次的汽车发售质量，聚焦于消费者诟病的负面情绪特征，加大研发投入，出产消费者满意的高性价比产品，能够很大程度上转变消费者目前对小米汽车舆情的分化形势。

②受于消费者对汽车性能配置、定价策略的聚焦关注，小米汽车品牌应保持对高质量车辆的研发，重视汽车性能配置，优化大众诟病的无用设施，融入消费者更需要且依赖的车辆配置，在智能化科技层方面保持自我创新。此外，对于汽车的定价策略，小米汽车可考虑独立出下级品牌，以研发符合多数消费者平价预期的产品，在一线高端轿车以及平价预期产品之间做出一定权衡，以赢得更佳的企业口碑。

（四）研究不足

本文研究总体上完成了社交媒体视角下对用户情感倾向性的分析探究，但受于资源成本限制等原因，研究存在如下不足：

①本文采用 word2vec 模型进行词向量处理，该训练模型虽广泛应用于中英文混杂的文本数据中，但由于其分词传入的特点，其在对词义多样性，如“一词多义”等更复杂情感文本的内容处理上，其表现不如 BERT 等更为成熟的预训练模型。但限于运算资源的限制，本文未能采用这种方式进行更为准确的情感分类。此外，本文的数据集构建采用伪标签法，并通过伪标签法得到的数据对模型进行训练，且最终得到的情感分类数据集均来自于机器分类，因此在情感分类精度上存在天然局限。因此为得到最为严谨的研究结论，仍需要后续以人工审校的方式进一步验证本文成果。

②本文在对复杂的非结构化文本数据进行处理时，虽对表情等非文字数据，采用字符转换的方式进行处理，但由于近年来视频社交媒体中，出现越来越多的评论图片表示，该类图片很大程度上能够决定用户评论时的情感倾向性，以及评论的含义表达。本文由于视频媒体网站的限制，未能得到该类影响话题特征以及情感倾向的图片信息，因此本研究在自然语言处理上仍需进一步探索。

参考文献

[1] 柳位平, 朱艳辉, 栗春亮, 等. 中文基础情感词词典构建方法研究[J]. 计算机应用,2009, 29(10): 2875-2877.

[2] 王大伟, 周志玮, 曹红根. 基于 PCA-SVM 算法的酒店评论文本情感分析研究[J].收藏, 2019, 21.

[3] 唐慧丰,谭松波,程学旗.基于监督学习的中文情感分类技术比较研究[J].中文信息学报,2007(06):88-94+108.

[4] 彭敏, 汪清, 黄济民, 等. 基于情感分析技术的股票研究报告分类[J]. 武汉大学学报(理学版), 2015, 61(2): 124-130.

[5] 胡朝举, 赵晓伟. 基于词向量技术和混合神经网络的情感分析[J]. 计算机应用研究,2018, 12.

[6] 陈葛恒. 基于极性转移和双向 LSTM 的文本情感分析[J]. 信息技术, 2018, 2.

[7] 王友卫,朱晨,朱建明,等.基于用户兴趣词典和 LSTM 的个性化情感分类方法[J].计算机科学,2021,48(S2):251-257.

[8] 韩坤,潘宏鹏,刘忠轶.融合 BERT 多层次特征的短视频网络舆情情感分析研究[J].计算机科学与探索,2024,(4): 1010-1020.

[9] Taboada M, Brooke J, Tofiloski M, et al. Lexicon-based methods for sentiment analysis[J].Computational linguistics, 2011, 37(2): 267-307.

[10] Kim S M, Hovy E. Automatic identification of pro and con reasons in online reviews[C]//Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions. 2006: 483-490.

[11] Hu M, Liu B. Mining and summarizing customer reviews[C]//Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. 2004: 168-177.

[12] Agarwal B, Mittal N. Semantic orientation-based approach for sentiment analysis[M]//Prominent feature extraction for sentiment analysis. Springer, Cham, 2016: 77-88.

[13] Ahmed M, Chen Q, Li Z. Constructing Domain-Dependent Sentiment Dictionary for Sentiment Analysis[J]. Neural Computing and Applications, 2020, 32(18):14719-14732.

[14] Zhang S, Wei Z, Wang Y, et al. Sentiment analysis of Chinese micro-blog text based on extended sentiment dictionary[J]. Future Generation Computer Systems,2018, 81: 395-403.

[15] Pang B, Lee L, Vaithyanathan S. Thumbs up? Sentiment classification using machine learning techniques[J]. arXiv preprint cs/0205070, 2002.

[16] Neethu M S, Rajasree R. Sentiment analysis in twitter using machine learning techniques[C]//2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT). IEEE, 2013: 1-5.

[17] Andreevskaia A, Bergler S. When specialists and generalists work together: Overcoming domain dependence in sentiment tagging[C]//Proceedings of ACL-08: HLT. 2008: 290-298.

[18] Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality[J]. arXiv preprint arXiv:1310.4546, 2013.

[19] Denil M, Demiraj A, Kalchbrenner N, et al. Modelling, visualising and summarizing documents with a single convolutional neural network[J]. arXiv preprint arXiv:1406.3830, 2014.

[20] Lan Z, Chen M, Goodman S, et al. Albert: A lite bert for self-supervised learning of language representations[J]. arXiv preprint arXiv:1909.11942, 2019

[21] Zhilin Yang;Zihang Dai;Yiming Yang;Jaime Carbonell;Ruslan Salakhutdinov;Quoc V. Le.XLNet: Generalized Autoregressive Pretraining for Language Understanding[J].2020,

[22] Chikersal P,Poria S,Ca Mbria E.SeNTU:Sentiment Analysis of Tweets by Combining a Rule-based Classifier with Supervised Learning[C]//SemEval@NAACL.HLT.2015:647.651.

[23] Kalchbrenner N, Grefenstette E, Blunsom P. A convolutional neural network for modelling sentences[J]. arXiv preprint arXiv:1404.2188, 2014.

[24] Chiorrini A,Diamantini C,Mircoli A,et a1.Emotion and sentiment analysis of twe ets using BERT[C]H EDBT/ICDT Workshops.2021.

[25] Zhengyan Zhang;Xu Han;Zhiyuan Liu;Xin Jiang;Maosong Sun;Qun Liu.ERNIE: Enhanced Language Representation with Informative Entities[J].2019.